DOCUMENT RESUME

ED 334 215 TM 016 665

AUTHOR Benton, Roberta L.
TITLE The Redundancy Index in Canonical Correlation Analysis.
PUB DATE Jan 91
NOTE 23p.; Paper presented at the Annual Meeting of the Southwest Educational Research Association (San Antonio, TX, January 24-26, 1991).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Correlation; Mathematical Models; *Multivariate Analysis; *Predictive Measurement
IDENTIFIERS *Canonical Redundancy Statistic; Redundancy Analysis; *Variance (Statistical)

ABSTRACT

The redundancy statistic (Rd) is discussed in relation to canonical correlation analysis. The index is a measure of the variance of one set of variables predicted from the linear combination of the other set of variables. A small data set (N=6) from the work of D. Clark (1975) was analyzed using SPSS-X. Two sets of two variables each were included. The variables of one set were considered criterion variables, and the variables of the second set were considered predictors. The CANBAK computer program was used to obtain squared structure coefficients, communality coefficients, and variate adequacy coefficients. The results suggest that the Rd was not truly multivariate and may have more value in the context of redundancy analysis than in canonical correlation. Rd is more a measure of prediction than association. It is an estimate of the shared variance of linear composites, not the variables themselves. The Rd is analogous to the squared multiple R in multiple regression. Redundancy coefficients computed in a canonical analysis may have somewhat limited value. Four tables present illustrative data. A 29-item list of references is included. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED334215

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

ROBERTA L. BENTON

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

The Redundancy Index

in

Canonical Correlation Analysis

Roberta L. Benton

Louisiana State University Medical Center

Paper presented at the annual meeting of the Southwest Educational Research Association, San Antonio, TX, January 25, 1991.

TM 016615

BEST COPY AVAILABLE

ABSTRACT

The redundancy statistic (Rd) is discussed in relation to canonical correlation analysis, employing a small data set for illustration. The index is a measure of the variance of one set of variables predicted from the linear combination of the other set of variables. Findings suggest that Rd is not truly multivariate and may have more value in the context of redundancy analysis than in canonical correlation.

The redundancy statistic in canonical correlation analysis has enjoyed varying degrees of acclaim and disfavor since its derivation in 1968. Some researchers laud the value of the redundancy coefficient in interpreting canonical results; whereas others note that the statistic does not simultaneously consider the intercorrelations of all variables and is multivariate only in that it involves several variables. The purpose of the present paper is to describe the redundancy coefficient and its function in canonical correlation. The historical context of its development and subsequent emergence of redundancy analysis are noted. Analysis of a small data set illustrates the concepts addressed.

Canonical Correlation Analysis

Canonical correlation analyzes the interrelationships of two sets of two or more variables. The method essentially simplifies the correlations between the domains by reducing the two sets of variables to a smaller set of weighted sums of the original variables, called canonical variates (Bentler & Huba, 1982). "The weights for the canonical variates represent a sort of compromise, under the constraint of orthogonality, between maximizing the interset covariance and minimizing the intraset variance" (Bock, 1975, p. 393). These two or more principle composites are utilized in lieu of individual variables for canonical correlation analysis (Hotelling, 1935).

The square of the canonical correlation (R_c^2) indicates variance shared by the canonical variates. As Pedhazur (1982)

notes, "It is very important to bear in mind that R_c^2 is not an estimation of the variance shared between X_p and X_q but of the linear combinations of these variables" (p. 722). The maximum value of R^2 remains the same whether we specify one set of variables predictor and the other criterion, or whether we exchange the roles (Hotelling, 1935).

Since canonical correlation concerns the degree of relationship between linear composites of two sets of variables and not the correlations between sets of variables, interpretation can become a problem (Lambert & Durand, 1975; Pedhazur, 1982; Stewart & Love, 1968). The canonical correlation, or R_c, designates the strength of the relationship between pairs of variates. When squared, R_c explains the variance in one canonical variate attributable to the other canonical variate. But as Stewart and Love (1968) observe, "It is important to note that a relatively strong canonical correlation may obtain between two linear functions, even though these linear functions may not extract significant portions of variance from their respective batteries" (p. 160). The redundancy statistic was proposed as a measure of interbattery relationship designed specifically to address this issue of interpretation within canonical correlation.

Lambert and Durand (1975) explain that the canonical R_c's measure the correlation between canonical scores (canonical variates) and do not measure the correlation between the sets of variables. Variance shared by linear composites does not reflect

the variance shared by the original independent and dependent variables. The canonical correlations may be numerically large and statistically significant, even though original data sets may share minimal variance (Lambert & Durand, 1975). Pedhazur (1982) further cautions against sole reliance on R_c^2 in interpretation of scientific findings.

In summary, difficulty in interpretation of canonical variate analysis is widely recognized (Cliff & Krus, 1976; Pedhazur, 1982; Thompson, 1984). Usefulness of canonical correlation depends not only on the degree to which canonical variates can summarize statistical association but also on "how well the variates can be conceptualized and interpreted," as Bentler and Huba (1982, p. 21) suggest. The redundancy index represents one alternative for evaluating statistical relationships between sets of variables.

Coefficients Associated with the Redundancy Index

Definitions of coefficients directly related to the redundancy index are listed as follows:

1. Structure coefficient. Correlation of an observed variable with a latent variable (Thompson & Borrello, 1985). In the context of this discussion structure coefficients are the correlation between an original variable and a canonical variate.
2. Squared canonical structure coefficient. "Proportion of variance linearly shared by a variable with the variable's canonical composite" (Thompson, 1984, p. 21).
3. Communality coefficient. "The sum of all of a variable's

squared structure coefficients" (Thompson, 1984, p. 25); it explains the usefulness of each variable in defining the canonical solution.

4. Variate adequacy coefficient. "The average of all the squared structure coefficients for the variables in one set with respect to one function" (Thompson, 1984, p. 25); it explains how well variate scores represent variance in unweighted variables.

5. Redundancy coefficient (Rd). Squared canonical correlation (Rc^2) times the average of all the squared structure coefficients for the variables in one set with respect to one function (Rc^2 X variate adequacy coefficient).

The Redundancy Statistic

The redundancy index [$\overline{R^2}$] in canonical correlation was derived by Stewart and Love (1968), and independently by Miller in 1969 (cited in Miller, 1975). The redundancy coefficient was proposed as a measure of the proportion of "the variance of C (one set of variables) predictable from P (the other set of variables), or the redundancy in C given P" (Stewart & Love, 1968, p. 161). The measure is non symmetric "since it would be rare indeed to find equal (average) predictability in both directions for two sets of variables" (Cramer & Nicewander, 1979, p. 43). The characteristic of non symmetry is desirable "because one set of variables may be almost completely subsumed by a larger set" (Stewart & Love, 1968, p. 160). The concept is represented graphically by Stewart & Love (1968) in the following diagram

Fig. 1. Nonsymmetric redundancy

in which redundancy is represented as the intersection of variable sets A and B. The redundancy index distinguishes between predictor and criterion variables, whereas features of canonical analysis are symmetric in regard to the sets of variables being analyzed (Thompson, 1984).

The redundancy index (Rd) measures the proportion of the variance in one set of variables predicted from the other set of variables. Thompson (1984) explains the derivation of the redundancy coefficient (Rd). To compute the statistic, the squared structure coefficients on a function are summed. The sum is divided by the number of variables in the set, yielding the variate adequacy coefficient for that set of variables with respect to one function. The variate adequacy coefficient for the given variable set is multiplied by the squared Rc for the function, yielding the redundancy coefficient. Rd is an estimate of the shared variance of linear composites, not the variance of the variables themselves. Stewart and Love derived Rd as a summary index, not as an analytic tool.

Subsequent to Stewart and Love's (1968) publication, Miller and Farr (1971) introduced Miller's comparable measure of interbattery relationship [R^2 y | x]. Miller (1975) went on to

develop a test for the significance of the bimultivariate redundancy statistic. Further, Miller (1975) noted, "Moreover, since it has also been shown that the 'general canonical correlation index' developed by Stewart and Love (1968) is, in fact, a special case of $R^2 y \mid x$, the significance test described here is applicable, without modification, to their canonical index" (p. 241). Cooley and Lohnes (1971) surmise that these separate developments have provided "a very useful expression for the degrees of relationship between batteries as displayed by the canonical model" (p. 171).

Gleason's (1977) paper offers the mathematical derivation for the redundancy index and describes some of the properties of the statistic. Levine (1977) cites previous approaches to between-set overlap, which are beyond the scope of this paper. Darlington, Weinberg, and Walberg (1973) suggest other measures of redundancy, and propose new methods for calculating similarity or overlap between two sets of variables. Cramer and Nicewander (1979) make a distinction between "redundancy measurement" and "measures of multivariate association" (p. 43) and enlarge upon the latter--theirs is the seminal treatment of Rd and its severe limits.

van den Wollenberg (1977) proposed a component method of extracting factors which maximizes the redundancy index of Stewart and Love (1968). He explained that canonical correlation only maximizes one part of the redundancy formula, i.e., Rc. He illustrated in a comparison of canonical correlation analysis and

what he terms redundancy analysis his method for maximizing redundancy. Muller (1981) explains that "Van den Wollenberg derived sets of variates which directly maximize the between set redundancy (whereas the canonical solution maximizes the between set canonical correlation)" (p. 139).

DeSarbo (1981) followed with canonical/redundancy factoring analysis for occasions where both correlation and redundancy are important, because "canonical correlation renders maximally related linear composites which, however, may not be very much related to the original sets of variables. Conversely, redundancy maximization renders linear composites maximally related to the original sets of variables, but may be somewhat unrelated to each other" (p. 328). For further application of redundancy analysis and related issues see Alpert and Peterson, 1972; Cramer and Nicewander, 1979; Darlington, Weinberg, and Walberg, 1973; Dawson-Saunders, 1982; DeSarbo, 1981; Lambert and Durand, 1975; Muller, 1981, or Tyler, 1982.

Example Illustrating the Redundancy Index

A small data set (see Table 1) from Clark (1975) was analyzed using SPSS-X. The sample (N=6) employed is unrealistic except for illustrative purposes, but provides heuristic value in that many of the mathematical computations can be derived by hand calculation for clarity in understanding the relationships inherent in canonical correlation. The example includes two sets of two variables each. The variables of one set are considered criterion variables and the variables of the second set are

predictors. The roles could be reversed or unspecified without affecting any of the results, i.e., the designations are arbitrary. The computer program CANBAK (Thompson, 1982) was employed for obtaining squared structure coefficients, communality coefficients, and variate adequacy coefficients.

INSERT TABLE 1 ABOUT HERE

Table 2 presents the characteristic roots and canonical correlations. The number of functions derived cannot exceed the number of variables of the smaller set. The first function is derived so as to explain the highest degree of intercorrelation between the variable sets. The first of the dimensions is predominant with an eigenvalue of 1.834. Root 1 (the first pair of canonical variates) accounts for 65% of the shared variance (0.64713); and root 2 accounts for approximately 2.5% (0.02578) of the remaining variance extracted.

INSERT TABLE 2 ABOUT HERE

Wilk's lambda is not statistically significant for either likelihood ratio. Sig of F [p calc] is greater than .05. In canonical analysis this test of statistical significance is used to determine statistically significant relationships among sets of variables. Wilk's lambda is an inverse criterion, i.e., the smaller the value of lambda, the more evidence there is for association effects (Cliff, 1987; Pedhazur, 1982; Stevens, 1986; Tatsuoka, 1971). The value of lambda ranges from zero to one. In multivariate associations lambda tests the relationship of the error sum of squares cross products (SSCP) matrix to the total

sum of squares cross products matrix. Tatsuoka (1971) differentiates between the simplest case of multivariate lambda (within groups SSCP matrix / total-sample SSCP matrix) and the application of lambda in the context of canonical analysis, where "the error SSCP matrix is the residual SSCP matrix after the effects of the correlations between the canonical-variate pairs have been removed" (p. 188). The larger the degree of association among variable sets in canonical correlation, the smaller the value of lambda. Lambda can also be described as a "multivariate omnibus effect size" (Thompson, 1990, p. 23) related to one minus R^2, the coefficient of alienation, with smaller values of lambda denoting larger effect sizes in relation to all canonical correlations in a study.

Since statistical significance is largely a function of sample size (Thompson, 1989), the reader should not be alarmed with these results. On the other hand, large sample sizes can lead to statistically significant results when result importance is negligible. We could obtain a statistically significant canonical correlation which could be "trivial from a substantive point of view" (Thompson, 1984, p. 20).

Structure coefficients, squared structure coefficients, communality coefficients and variate adequacy coefficients for these data are presented in Table 3. Using Pedhazur's (1982) rule of thumb that structure coefficients greater than or equal to 0.30 be treated as meaningful, one coefficient would fall below that standard (0.219). The variate adequacy coefficients

indicate that 57% of the total variance of the first variable set is extracted by the first canonical variate; and that 43% of the total variance of the first set was extracted by the second canonical variate. For the second variable set the corresponding values are 53% and 47%, respectively. Communality coefficients are all 1.00. With two variables per set we would expect all of the variance to be accounted for; but if one variable set had been larger, the two possible functions may not have extracted everything from the larger set, causing a communality coefficient to be less than one.

INSERT TABLE 3 ABOUT HERE

Redundancy coefficients can be consulted for guidance in interpreting several questions in canonical results (Thompson, 1984): (a) "To what extent can one set of two or more variables be predicted or explained by another set of two or more variables?" (p. 60), and (b) "What relative power do different canonical functions have to predict or explain relationships?" (p. 62).

Table 4 presents the redundancy coefficients. Redundancy coefficients for the dependent variable set indicate that 36.73% of the total variance of the first variable set (criterion) is predicted from or explained by the first canonical variate of the second set (predictor), and 1.12% of the shared variance is explained by the second canonical variate. The pooled redundancy coefficient for the criterion variable set was 37.85% (36.73% + 1.12%). For the predictor variable set (covariates) the values

are 34.30% and 1.21%, respectively, with a total redundancy coefficient of 35.51%. The total explanatory power of the covariates to explain variance shared is 35.51%. The values are similar across the two variable sets in the present example but can be radically different in reality.

INSERT TABLE 4 ABOUT HERE

The canonical redundancy coefficient has proven useful in determining whether results of canonical analysis consist of a "g"-function, explained by Thompson (1990) as a function with a very large Rc and very large structure coefficients on all or most variables. However, other uses are dubious.

Example Illustrating that Squared Multiple R and Rd are Analogous i.e., not Multivariate

In canonical redundancy analysis "...the explained variance of the Y-variables is equal to their respective squared multiple correlation with the X-variables. This is also true for canonical correlation analysis of which multiple correlation is a special case" (van den Wollenberg 1977, p. 211).

To examine the concept of redundancy in the context of multiple regression, two multiple regression analyses were done using both predictors with each criterion variable separately. The average of the two separate multiple R^2 was computed as follows: Sq. Mul. R for Pred 1, Pred 2 with Crit 1 = 0.26247. Sq. Mul. R for Pred 1, Pred 2 with Crit 2 = 0.49448. 0.26247 + 0.49448 = 0.75695 / 2 = 0.378475 x 100 = 37.85%. The result (37.85%) is exactly the same as the 37.85% result for cumulative

dependent redundancy. The total redundancy of the first variable set given the second variable set equals the average of the squared multiple correlations of each criterion variable with all of the predictor variables (Pedhazur, 1982; Stewart & Love, 1968). This property makes redundancy "synonymous with average predictability" (Cramer and Nicewander, 1979, p. 43). Thus, redundancy coefficients are averaged univariate statistics (Thompson, 1990), and are not truly multivariate.

Conclusions

Three conclusions are important to the researcher and the educator and should be noted. First, Rd measures the predictable variance of one set of variables from the linear composite of the other set of variables, and hence is more a measure of prediction than association (Muller, 1981). It is an estimate of the shared variance of linear composites, not the variables themselves. The redundancy statistic was developed as a summary index, not an analytic tool (Stewart & Love, 1968); albeit the function of the index has been enlarged given the development of redundancy analysis and related techniques.

Second, the redundancy statistic is analogous to the squared multiple R in multiple regression. For this reason the measure is not truly multivariate. It does not take into simultaneous consideration all the intercorrelations of the predicted variables. The redundancy index is multivariate only in that it involves several (but not all) variables at a time (Cramer & Nicewander, 1979).

Third, "Redundancy coefficients computed in a canonical analysis may have somewhat limited value, [in that] the statistic seems to make more sense in the context of redundancy analysis or some variant of redundancy analysis" (Thompson, 1984, pp.29-30). It does not make sense to emphasize the interpretation of results not optimized as part of the analysis. Rc is optimized in canonical analysis; Rd is not (Thompson, 1984).

References

Alpert, M. I. & Peterson, R. A. (1972). On the interpretation of canonical analysis. Journal of Marketing Research, 9, 187-192.

Bentler, P. M. & Huba, G.J. (1982). Symmetric and asymmetric rotations in canonical correlational analysis: New methods with drug variable examples. In Hirschberg, N. & Humphreys, L. G., (Eds.), Multivariate applications in the social sciences (pp. 21-46). Hillsdale, NJ: Lawrence Erlbaum Associates.

Bock, R. D. (1975). Multivariate statistical methods in behavioral research. New York: McGraw-Hill.

Clark, D. (1975). Understanding canonical correlation analysis. Norwich: Geo Abstracts, University of East Anglia.

Cliff, N. (1987). Analyzing multivariate data. San Diego: Harcourt Brace Jovanovich.

Cliff, N. & Krus, D. J. (1976). Interpretation of canonical analysis: Rotated vs. unrotated solutions. Psychometrika, 41(1), 35-42.

Cooley W. W. & Lohnes, P. R. (1971). Multivariate data analysis. New York: John Wiley.

Cramer, E. M. & Nicewander, W. A. (1979). Some symmetric, invariant measures of multivariate association. Psychometrika, 44(1), 43-54.

Dawson-Saunders, B. K. (1982). Correcting for bias in the

canonical redundancy statistic. Educational and Psychological Measurement, 42, 131-143.

DeSarbo, W. S. (1981). Canonical/redundancy factoring analysis. Psychometrika, 46(3), 307-329.

Darlington, R. B., Weinberg, S. L. & Walberg, H. J. (1973). Canonical variate analysis and related techniques. Review of Educational Research, 43(4), 433-454.

Gleason, T. C. (1976). On redundancy in canonical analysis. Psychological Bulletin, 83(6), 1004-1006.

Hotelling, H. (1935). The most predictable criterion. The Journal of Educational Psychology, 26, 139-142.

Lambert, Z. V. & Durand, R. M. (1975). Some precautions in using canonical analysis. Journal of Marketing Research, 12, 468-475.

Levine, M. S. (1977). Canonical analysis and factor comparison. Beverly Hills: Sage.

Miller, J. K. (1975). The sampling distribution and a test for the significance of the bimultivariate redundancy statistic: A monte carlo study. Multivariate Behavioral Research, 10, 233-244.

Miller, J. K. & Farr, S. D. (1971). Bimultivariate redundancy: A comprehensive measure of interbattery relationship. Multivariate Behavioral Research, 6, 313- 324.

Muller, K. E. (1981). Relationships between redundancy analysis, canonical correlation, and multivariate regression. Psychometrika, 46(2), 139-142.

Pedhazur, E. J. (1982). Multiple regression in behavioral research (2nd ed.). New York: Holt, Rinehart and Winston.

Stevens, J. (1986). Applied multivariate statistics for the social sciences. Hillsdale, NJ: Lawrence Erlbaum Associates.

Stewart, D. & Love, W. (1968). A general canonical correlation index. Psychological Bulletin, 70(3), 160-163.

Tatsuoka, M. M. (1971). Multivariate analysis: Techniques for educational and psychological research. New York: John Wiley.

Thompson, B. (1982). CANBAK: A computer program which performs stepwise canonical correlation analysis. Educational and Psychological Measurement, 42, 849-851.

Thompson, B. (1984). Canonical correlation analysis: Uses and interpretation. Newbury Park: Sage.

Thompson, B. (1989). Statistical significance, result importance, and result generalizability: Three noteworthy but somewhat different issues. Measurement and Evaluation in Counseling and Development, 22(4), 2-6.

Thompson, B. (1990). Finding a correction for the sampling error in multivariate measures of relationship: A Monte Carlo study. Educational and Psychological Measurement, 50(1), 15-31.

Thompson, B. & Borrello, G. M. (1985). The importance of structure coefficients in regression research. Educational and Psychological Measurement, 45, 203-209.

Tyler, D. E. (1982). On the optimality of the simultaneous redundancy transformations. Psychometrika, 47(1), 77-86.

van den Wollenberg, A. L. (1977). Redundancy analysis an alternative for canonical correlation analysis. Psychometrika, 42(2), 207-219.

Table 1

Data Set (Clark, 1975)

	CRIT1	CRIT2	PRED1	PRED2
1	07	18	37	12
2	11	13	18	40
3	21	07	16	12
4	20	15	09	11
5	15	15	14	14
6	11	01	01	11

Table 2

Eigenvalues and Canonical Correlations

Root	Eigenvalue	Pct.	Cum. Pct.	Canon Cor.	Sq. Cor.
1	1.83394	98.58	98.58	0.80445	0.64713
2	0.02647	1.42	100.00	0.16057	0.02578

Dimension Reduction Analysis

Roots	Wilk's L.	F	Hyp df	Error df	Sig of F
1 to 2	0.34377	0.70556	4	4	0.628
2 to 2	0.97422	0.07940	1	3	0.796

Table 3

Correlations between Dependent and Canonical Variables

	1	2
CRIT1	-.61719	.78682
CRIT2	.86851	.49567

Squared Structure Coefficients

CRIT1	.381	.619
CRIT2	.754	.246

Communality Coefficients

CRIT1	1.000
CRIT2	1.000

Variate Adequacy Coefficients

.568	.432

Correlations between Covariates and Canonical Variables

	1	2
PRED1	.97576	.21883
PRED2	.32859	-.94447

Squared Structure Coefficients

PRED1	.952	.048
PRED2	.108	.892

Communality Coefficients

PRED1	1.000
PRED2	1.000

Variate Adequacy Coefficients

.530	.470

Table 4

Variance expl by canonical var of Dependent Variables

Can Var	%Var DEP	Cum % DEP	% Var COV	Cum % COV
1	56.76	56.76	36.73244	36.73244
2	43.29	100.00	1.11483	37.84727

Variance expl by canonical var of Covariates

Can Var	%Var DEP	Cum % DEP	% Var COV	Cum % COV
1	34.30092	34.30092	53.00433	53.00433
2	1.21171	35.51262	46.99567	100.00000